Turkish Journal of Forecasting vol. 9 no. 1 (2025) pp 18-29

Content list available at JournalPark

Turkish Journal of Forecasting

Journal Homepage: tjforecasting.com

A Resarch on the Factors Affecting the Outcomes of Child Abuse Cases Using Machine Learning Methods

Saime Şule Aksakal[1,*], Erol Eğrioğlu[2]

[1]*Department of Mathematics, Faculty of Arts and Sciences, Giresun University, Gure Campus, 28200 Giresun, Turkey*
[2]*Department of Statistics, Faculty of Arts and Sciences, Giresun University, Gure Campus, 28200 Giresun, Turkey*

Abstract

Modern information technology makes it possible to collect and store scientific and social research data. Some statistical methods can provide quite reliable results when the necessary assumptions are met in uncovering existing or hidden relationships between data. However, since data collected from real life often do not meet these assumptions, data mining methods that require fewer assumptions and can be applied to flexible and complex data sets have been developed for prediction. The use of machine learning methods, which include data mining techniques, to process data and produce meaningful information has become widespread in recent years. In this study, techniques such as the CHAID algorithm, an application of decision trees, and support vector machines, were compared with the logistic regression analysis method. The study's sample consists of data from 61 child abuse cases in which the UCIM Saadet Öğretmen Association Struggling Child Abuse requested participation. The dependent variable of the study is whether the defendant received a sentence at the end of the trial, while the independent variables are five variables identified by leveraging expert (lawyer) opinions. As a result, it was found that the CHAID algorithm and support vector machines provided more accurate classification.

Keywords: Machine Learning, Logistic Regression, Support Vector Machines, CHAID Algorithm, Child Abuse Cases.

1. Introduction

It is now possible to collate and store scientific research data using the latest information technology. The issue of how to deal with the future-oriented estimation methods that would be rendered superfluous by an increase in the number of data items has been resolved by the development of data mining techniques (Acuna & Rodriguez, 2004). The purpose of this method is to identify the functions that will enable the prediction of large data sets (Fouché & Langit, 2011). The objective of the method is to provide an alternative to classical statistical techniques by employing advanced computer techniques to address the same problems (Castro et al., 2007). Data mining represents the principal subfield of machine learning that is concerned with transforming large data repositories into useful information (Chen et al., 1996). The process of transformation is comprised of several stages, including data pre-processing, data transformation, data integration, data reduction, application, and presentation (Maharana et al., 2022). In the process of data transformation, artificial intelligence, statistics (relating data sets to numerical relationships) and machine learning (learning from data sets to make predictions) are utilised (Raschka et al., 2020). In the

* Corresponding Author.
E-mail addresses: sule.aksakal@giresun.edu.tr (Saime Şule Aksakal), erol.egrioglu@giresun.edu.tr (Erol Eğrioğlu)

ORCID ID:
Saime Şule Aksakal : 0000-0002-1810-1040
Erol Eğrioğlu : 0000-0003-4301-4149

context of data mining, if the target is clearly defined, supervised learning is employed; conversely, unsupervised learning is utilised when the target is uncertain (Chapman et al., 2000).

In this study, the increasing number of cases of child abuse in recent years has been taken into account, as well as the social response to this phenomenon (Fan et al., 2021). In order to address this issue, decision trees, support vector machines and logistic regression methods have been employed (Wang & Ding, 2020). The reason for choosing these techniques is both their high classification performance and the ability for users to define the models' hyperparameters. During the study, multiple machine learning models were applied to the dataset, and the most suitable parameters were determined.

Child abuse is defined as the intentional infliction of physical, emotional or sexual harm upon a child (Lippard & Nemeroff, 2020). In accordance with the definition provided by the World Health Organization (WHO), child sexual abuse is defined as the exploitation of a child for the purpose of sexual gratification or manipulation by a parent, guardian, or adult (Mathews & Collin-Vézina, 2019). Such maltreatment has been demonstrated to cause profound and long-lasting harm to the child in question, from a physiological, psychological and emotional perspective (Chadaga et al., 2024). The perpetrators of sexual violence against children, prostitution, child pornography and child sexual exploitation are causing harm to children (Walker-Descartes et al., 2021). The World Health Organization (WHO) has estimated that at least 150 million girls and 73 million boys have been subjected to forced sexual intercourse or physical contact involving any form of sexual abuse (WHO, 2002). The majority of children who have been subjected to sexual abuse are reluctant to disclose the abuse for various reasons (Paine & Hansen, 2002). The inability to articulate the abuse due to the victim's young age, the perpetrator's use of intimidation (Allen-Collinson, 2009), the fear of the victim being rejected by their family or carers, and the abuse being motivated by incestuous desires are all factors that contribute to the difficulty in recognising the abuse (Bowlby, 1984). The reluctance to compromise the integrity of the narrative and the associated apprehension regarding the lack of belief and support from the individual in question can impede the ability of children to communicate (Kassin & Gudjonsson, 2004). In particular, domestic violence and the tendency for such incidents to be concealed due to societal norms are significant factors contributing to the underreporting of abuse cases involving male victims (Zalcberg, 2017). In the event that victims of sexual abuse, or their proxies, report the incident to the relevant authorities, the process will be subject to the jurisdiction of the state(Kruttschnitt et al., 2014). Child observation centres, police, gendarmerie and public prosecutor's offices serve as points of contact for those wishing to make a report. In the judicial phase, the state and civil society organisations engaged in child labour initiatives provide voluntary legal and psychological assistance to children (Bajpai, 2018). In Turkey, the leading civil society organisation providing support to children and families affected by child abuse is the Saadet Öğretmen Çocuk İstismarı ile Mücadele Derneği (UCİM Saadet Öğretmen Association Struggling Child Abuse). In 2020, UCIM established the first European office for the prevention of child neglect and abuse in Izmir, guided by its mission to protect children's rights. Currently, the prevention offices in 13 of Turkey's provinces are engaged in active work, providing educational, legal and rehabilitative assistance, and implementing various projects in their respective fields (UCIM, 2024). Furthermore, it maintains active projects and coordinators in 50 regions, continuing its advocacy for children's rights. In addition to the voluntary mental health professionals who provide rehabilitation support, the organisation also employs volunteer lawyers who offer legal assistance to child victims of abuse in the context of the legal process. The cases received via the UCIM helpline are evaluated and legally pursued by temporary attorneys. This work, which requires the acquisition of necessary ethical permissions, provides an illustrative example of the 61 cases of child sexual abuse that UCIM has monitored.

2. Method And Procedure

2.1. Decision Trees

The decision tree structure is frequently selected for its ease of construction and comprehensibility, making it a preferred method for classification and regression models (Huysmans et al., 2011; Yücesoy et al., 2023). The structure of decision trees for non-parametric statistical data, which assumes a normal distribution and homogeneous variance, is based on the "if-then" conditional relationships (Dumitrescu et al., 2022). The relationship between the dependent variable and the independent variables is classified according to the levels of the latter. The classification is created by displaying each variable on a node in the diagram. The tree begins with a root node representing all the examples and then proceeds to divide the data into subgroups by cutting the tree into branches (Liu & Cocea, 2019). The most frequently employed techniques in decision trees are AID, CHAID, CART (C&ART), and QUEST algorithms. In the AID algorithm, the dependent variable must be quantitative, while the independent variables must be categorical. A node may be divided into no more than two nodes. The quantitative dependent variable is therefore an example of a regression tree, with the splitting criterion being the total of inter-group squares. In the CHAID algorithm, the dependent variable must be quantitative or categorical, while the independent variable must be categorical. In contrast to the AID algorithm, the CHAID algorithm allows for the division of a node into two or more branches. The dependent variable may be regressed or classified using a tree structure, with the splitting criterion being either an F-test or a Chi-square test (Alp & Öz, 2019). In the CART algorithm, both dependent and independent variables can be quantitative or categorical, and, similarly to the AID algorithm, it allows for only two splits at each node. In the event that the dependent variable is categorical, the twoing or Gini method is employed; conversely, if the dependent variable is quantitative, the smallest area method is utilised. In the QUEST algorithm, the dependent variable must be categorical, while the independent variable can be either categorical or numerical. In the case of nodes, two divisions are permitted, and the quadratic discriminant criterion is employed (Alp & Öz, 2019).

2.2. CHAID Algorithm

The criterion of the CHAID algorithm, which is one of the decision tree techniques, is that the dependent variable can be quantitative or categorical, while the independent variable must be categorical (Akin et al., 2017). The CHAID technique is an improved version of the AID technique, which was specifically designed for categorical dependent variables (Kass, 1980). The chi-square test statistic, used as the splitting criterion, iteratively divides the existing dataset into subgroups using the 'if-then' condition and ranks the independent variables that affect the dependent variable in a tree-like structure based on their effect size.

The chi-square analysis is a non-parametric method for calculating the significance of the relationship between two categorical variables. This is achieved through the application of statistical techniques, specifically the construction of a contingency table, which allows for the examination of the relationship between the two categories. For a cross table of size $r \times c$, let $i = 1,2,\ldots,r$ be the row indices and $j = 1,2,\ldots,c$ be the column indices. G_{ij} represents the observed value in the i-th row and j-th column, while B_{ij} denotes the expected value of the ij-th observation. The χ^2 test statistic is obtained using the formula:

$$\chi^2{}_{test} = \frac{\sum_{j=1}^{c}\sum_{i=1}^{r}(G_{ij}-B_{ij})^2}{B_{ij}} \quad (1)$$

If $\chi^2{}_{test} > \chi^2{}_{(r-1)(c-1)}$, the difference between the expected and observed values is found to be significant, leading to the conclusion that there is a statistical relationship between the two variables. The procedures are conducted in a manner whereby the relationship between each independent variable and the dependent variable is determined, and the process is continued until all observations are homogeneous. Additionally, decision tree branching criteria are established (Alp & Öz, 2019).

A set of k independent variables $(k = 1,2,\ldots,l)$ and one dependent variable (Y) are defined. The dependent variable let us consider r categories $(i = 1,2,\ldots,r)$ of the dependent variable and c categories $(j = 1,2,\ldots,c)$ of the independent variable.

Step 1. If the relationship between X_k and Y is expressed in a 2xc-dimensional contingency table (where X_k is an independent variable comprising two subcategories), the procedure progresses to Step 4. Otherwise, the procedure progresses to Step 2.

Step 2: The objective is to establish a comprehensive two-dimensional contingency table comprising all possible combinations of X_k and Y, and to calculate the square root of the correlation coefficient. The objective is to construct $r > 2$ X_k's subordinate categories and generate $C(r,2) = \frac{r!}{2!(r-2)!}$ number of contingency tables.

Step 3: In the construction of contingency tables for X_k and Y, the condition of ensuring that the lowest value of the test statistic, denoted by $\chi^2{}_{(1)(c-1)}$ is met is fulfilled by selecting the table with the lowest $\chi^2{}_{test}$. The two subcategories pertaining to the X_k variable are consolidated into a unified category, which is then designated as such. The value of r is updated to r-1 and the process returns to Step 1. This step is repeated for all constructed bivariate contingency tables until the condition of $\chi^2{}_{test} > \chi^2{}_{(1)(c-1)}$ is satisfied. Once this condition has been met, the process advances to Step 4. The preceding steps 1-4 are applied to all independent variables $(X_1, X_2, \ldots, X_l)$.

Step 4: The $\chi^2{}_{test}$ values for the $2xc$-dimensional contingency tables comprising the variables $X_1\_Y,\ldots,X_k - Y,\ldots,X_l\_Y$ are calculated. The independent variable, X_k, which has a value of $\max(\chi^2{}_{test})$, is selected from the range of options available in the contingency table. The lower-level categories and combined categories are generated through the process of branching and sub-nodes. The steps of the algorithm are repeated for each child node. All observations at a node are dependent on the lower category of the dependent variable, or the $X_k - Y$ variable pair. In the event that the entire $2xc$ contingency table is found to be in accordance with the condition of $\chi^2{}_{test} < \chi^2{}_{(1)(c-1)}$, the terminal node is designated as the "terminal node." The observations made at the terminal node are homogeneous and cannot be classified into different categories. The process continues until all the nodes in the tree, except for the terminal node, have been considered.

2.3. Support Vector Machines (SVM)

The Support Vector Machine (SVM) algorithm was initially proposed by Vapnik in 1963 as a method for constructing a linear classifier (Vapnik, 1963). SVM represents a supervised learning model. The input data is labelled according to a specific class. SVM employs an n-dimensional hyperplane to distinguish between two classes of data provided as input (Noble, 2006). Hyperplanes are also referred to as decision boundaries. The decision boundary is constructed to be as distant as possible from the nearest data points of each class. The data points that define the hyperplane are referred to as "support vectors." In non-linear models, the distance between the two classes is calculated using a kernel, which is a mathematical structure with this specific purpose. A kernel is a function that enables the construction of high-dimensional and non-linear models. In the context of non-linear problems, the kernel function can be employed to augment the unprocessed data with additional dimensions, thereby transforming the unprocessed data into a linear problem within a higher-dimensional space. The kernel function has been designed to facilitate the expeditious completion of specific calculations.

The decision plane of the linear SVM is highly effective at distinguishing between classes. In other words, a distinctive boundary between the two classes can be identified for the two-class labelled data sets. As indicated in equation 2, for a known

(supervised learning) data set with labels, x represents the feature vector. The variable y_i represents the labels associated with the training data, which may be either +1 or -1. The value n denotes the total number of features present in the data set.

For $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), (x_i, y_i) \in \mathbb{R}^n \times \{-1, +1\}$ (2)

The optimal hyperdimensional thought process (3) is formulated in the optimal hyperdimensional thought process (3). In this context, w represents a hyperdimensional's normal vector, x_i represents an input feature vector, and b represents the bias value.

$$wx^T + b = 0$$

$$wx_i^T + b \geq +1 \; if \; y_i = +1 \quad (3)$$

$$wx_i^T + b \leq -1 \text{ if } y_i = -1$$

In accordance with the aforementioned equation (4), two parallel hyperplanes, designated as H_1 and H_2, can be expressed as follows:

$$H_1: wx_1^T + b = +1 \quad (4)$$

$$H_2: wx_2^T + b = -1$$

The difference between the H_1 and H_2 planes, obtained through the use of equation (4), represents the distance between the two planes.

$$wx_1^T + b = +1$$

$$wx_2^T + b = -1$$

$$= w(x_1^T - x_2^T) + b = +2$$

$$= \left(\frac{w}{\|w\|}(x_1^T - x_2^T) + b\right) = \frac{2}{\|w\|} = \frac{2}{\sqrt{w.w}} \quad (5)$$

It is proposed that the inter-plane distance, H_1 and H_2, is equal to $2\|w\|$, where w is a vector. Furthermore, it is assumed that the H_0 plane is situated at an equal distance from both H_1 and H_2. In this case, the H_0 plane can be represented by the following equation:

$$H_0 : wx_0^T + b = 0 \quad (6)$$

The distance between H_1 and H_2 planes is denoted by d^{positive}. Similarly, the distance between the H_0 ve H_2 planes is also represented by d^{negative}. In accordance with the calculations provided by equation (6), these distance values are $\frac{1}{\|w\|}$. The fundamental objective of training an SVM model is to calculate the values of w and b. This enables the optimal partitioning of high-dimensional data, thereby maximising the margin.

In the event that two classes can be linearly separated, as previously described, the nearest support points on the hyperplane are identified using distance measurements based on the support vector. The maximum distance is used to determine the hyperplane, which is then used to separate the two classes (Cheng et al., 2021). In the event that the two classes cannot be separated linearly, the support vectors attempt to identify the maximum hyperplane, taking into account the potential classification error (Pant & Kumar, 2022). The aforementioned non-linearity of the data can be resolved by incorporating the variables representing the error terms, denoted by ε_i, into the optimisation model. In non-linear support vector machines, the problem can be solved by transforming the data into a linear format using non-linear core functions (Aydav & Minz, 2020). The

most well-known core functions encountered in the literature are linear, radial, polynomial, and sigmoid core functions (Tso & Mather, 2009).

2.4. Logistic Regression Method

Logistic regression is a classification method employed to investigate the effect of a dependent variable comprising two or more categories on each of its independent variables. In circumstances where categorical variables have been classified according to a scaling system, logistic regression analysis represents a viable analytical technique. This is a regression technique whereby the expected values of the response variable are obtained in a probabilistic manner, with respect to the independent variables. The assumption of a normal distribution does not require the prior assumption of persistence. The aim is to determine the effects of the independent variables on the dependent variable. Logistic regression can be classified into three main categories.

Binary logistics regression is a statistical method used to analyse dependent variables that fall into two categories and independent variables that possess either a continuous or categorical attribute. The dependent variable is a measure of the outcome under investigation, with a value of 1 in a given context and 0 in other scenarios. The independent variables pertain to the units included in the groups identified through this type of modelling. The probability of her belonging to the first category of a unit is calculated using the following formula:

$$\mathrm{P}(Y_i = 1) = \frac{e^{(b_0+b_1x_{i1}+\ldots+b_kx_{ik})}}{1+e^{(b_0+b_1x_{i1}+\ldots+b_kx_{ik})}} \tag{7}$$

Ordered logistics regression model is used when the dependent variable contains more than two categories and these categories are ranked from smallest to largest. In the Y_i category, the probability value is represented by r, the edge value for the relevant category is represented by i, the regression coefficients are represented by b, and the vector of scaling parameters is represented by Z, which provides insight into the scaling parameters for the aforementioned categories. This probability value is calculated using the formula:

$$Link(Y_i)=\frac{r_i[b_0+b_1X_1+\ldots+b_kX_k]}{exp[\phi_0+\phi_1Z_1+\ldots+\phi_tZ_t]} \tag{8}$$

Multinomial logistic regression analysis should be used when the dependent variable has two or more categorical classification scales. For example, if there is a dependent variable consisting of students enrolled in three different academic programs, it can be analyzed using multinomial logistic regression analysis. This probability is calculated using the formula

$$\mathrm{P}(Y_i = n) = \frac{e^{(b_{n0}+bn_1X_{i1}+\ldots+b_{nk}X_{ik})}}{1+\sum_{n=1}^{M-1}e^{(b_0+b_1X_{i1}+\ldots+b_kX_{ik})}} \tag{9}$$

The Maximum Likelihood method, which is similar to the Ordinary Least Squares (OLS) method used in regression analysis for parameter estimation and model evaluation, is employed for models established for classification purposes. This method ensures that the coefficients of the independent variables are obtained in such a way as to maximize the probability of occurrence of the dependent variable, structured similarly to dummy variables. If the primary application area of the logistic regression model is defined as classification, the model's accuracy in classification can be examined using traditional classification performance metrics (Nguyen-Thihong & Vo-Van, 2024).

3. Applications

In this study, data from 61 reports of child sexual abuse received through the UCİM hotline, for which ethical permissions were obtained, were examined, focusing on the independent variables that could affect the defendant's sentencing at the end of the judicial process. The average age of the 61 children is 11, with the youngest being 2 and the oldest 18 years old. Among the children, 49 are girls (80.3%) and 12 are boys. Three of the children have special needs. The average age of the defendants is 48.5, and 89% of them are acquaintances of the child. Due to threats, 64% of the children were unable to report the abuse immediately. The number of children who were able to access justice in a timely manner after the abuse was noticed is 27. Approximately 63% of all cases involved qualified sexual abuse. During the judicial process, 63% of the defendants were held in custody. In 50 of the 61 cases (82%), UCİM's legal support for children and their families was deemed appropriate by the court. In 40 of all cases (65.6%), the defendants received sentences, while 21 (34.4%) were acquitted.

Before the models were applied, the datasets were split into training and test data in ratios of 80% and 20%, respectively. Grid search approach has been used to adjust the hyperparameters on the training dataset. In this study, a 10-fold cross-validation method repeated five times was used. The optimal hyperparameters for each algorithm were determined based on the training dataset, and test data, which had not been used in the algorithm previously, were employed to validate the models. For the tree-based methods, the CHAID algorithm utilized the information gain method to select the most effective features. Information gain measures how much each feature explains the target variable. This method uses the entropy value to measure irregularity. The complexity parameter selected through cross-validation is chosen from the search space {0, 0.01, . . ., 0.09, 0.1, 0.2, . . ., 0.9, 1.0. In the support vector machine, hyperparameter selections were calculated using repeated grid search and cross-validation

methods between the values Cost (C) = 1 and Gamma = 0.05. In the logistic regression model, the distribution parameter for the Binomial family was assumed to be 1 and significance code is chosen from the search space {0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1}.

Table 1. Dependent and independent variables obtained from case monitoring forms

Variable Type	Variable	Variable Category
Dependent Variable	Sentencing Status	Sentenced (Yes) / Not Sentenced (No)
Independent Variable	Child's Age*	0-6 Years / 7-11 Years / 12-18 Years
Independent Variable	Recurrence/Duration of Abuse	Single Instance of Abuse / Systematic Abuse
Independent Variable	Type of Sexual Abuse	Penetrative Sexual Abuse / Contact without Penetration
Independent Variable	Timely Access to Justice	Yes (Timely Access) / No (Delayed Access)
Independent Variable	UCİM's Participation in the Case	Yes (Participation Request Accepted) / No (Participation Request Not Accepted)

*Categorical age levels were determined through a literature review as 0-6 years, 7-11 years, and 12-18 years

The results of the sentencing status were analysed using the CHAID method, which identified the independent variables that had an impact on the outcome. These were the defendant's participation in the case, timely access to justice, the duration and recurrence of the abuse, the type of abuse and the age of the child. The results are presented in Table 2.

Table 2. Independent variables affecting the case outcome variable

Variable	Decision Tree Split	Test Statistic Value (χ^2)	P
UCİM's Participation in the Case	Binary Split	15,828	0,000
Timely Access to Justice	Binary Split	7,351	0,007
Recurrence/Duration of Abuse	Binary Split	3,891	0,049
Type of Abuse	Binary Split	5,844	0,016
Child's Age	Binary Split	4,950	0,026

The accuracy of the constructed decision tree was tested using the cross-validation technique, with a cross-validation fold value of 10. The risk indicators calculated are presented in Table 3.

Table 3. Risk indicators

Method	Estimate	Std. Error
Resubstitution	0,115	0,041
Cross-Validation	0,213	0,052

The classification matrix of the obtained decision tree was calculated using Table 4, and the accuracy value was determined to be 88.5%.

Table 4. Classification matrix of the decision tree

	Predicted		
Observed	Sentenced (Yes)	Not sentenced (No)	Percent correct
Sentenced (Yes)	35	5	0,875
Not sentenced (No)	2	19	0,905
Overall percentage	0,607	0,393	0,885

According to the decision tree (Fig.1) obtained using the CHAID algorithm the independent variables that most influence the defendant's sentencing status are, in order of impact, UCIM's involvement in the case, timely access to justice, the duration of abuse, the type of abuse, and the child's age. All independent variables from the root node to the leaves and their effect sizes on the outcome are listed in the following items:

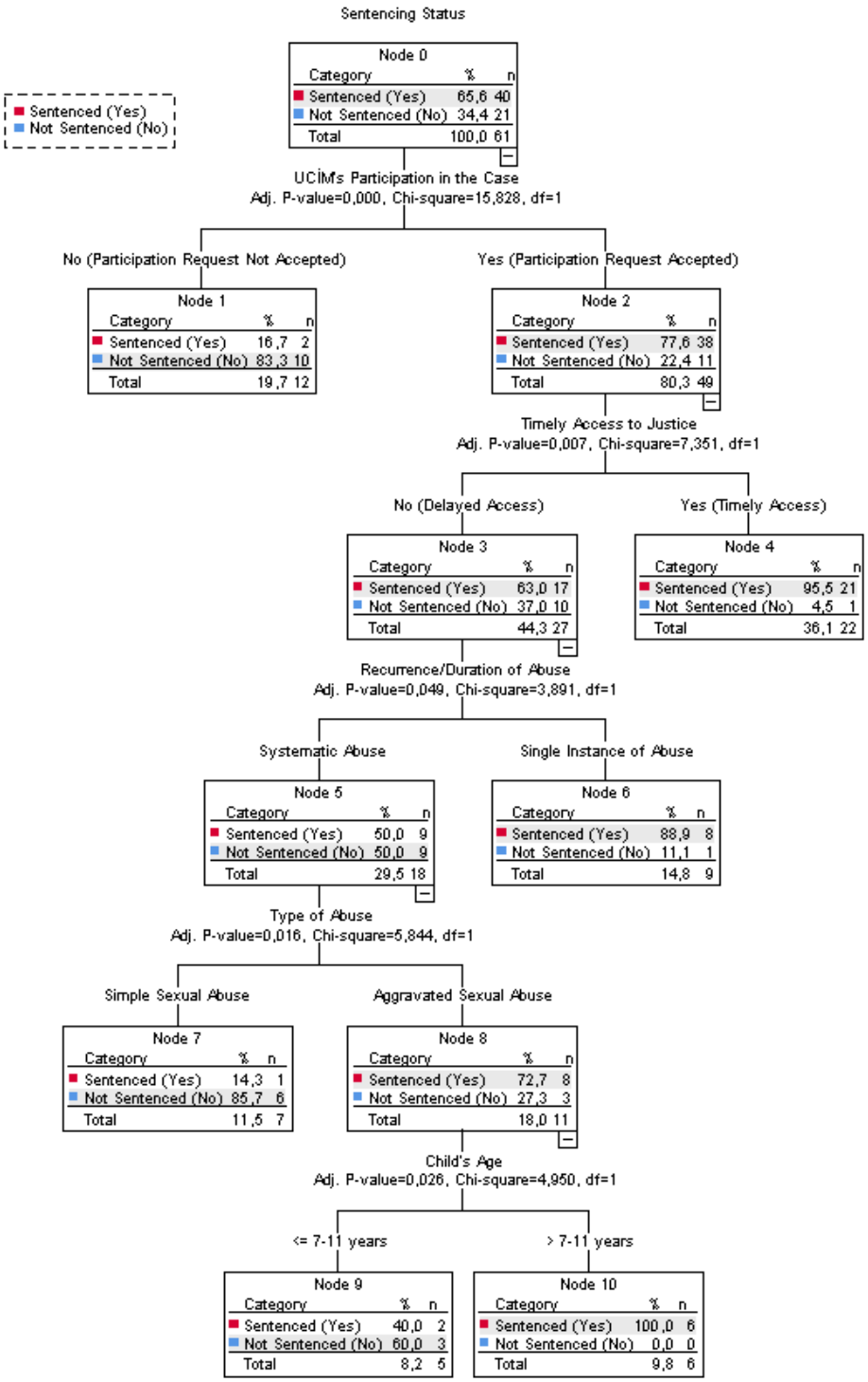

Figure 1. The decision tree obtained with the CHAID algorithm

- The entire set of cases was analyzed using the CHAID algorithm, and all decision rules of the algorithm were established with an "if-then" rule structure as detailed below: Out of all the cases, 40 cases (65.6%) resulted in the defendant receiving a sentence, while 21 cases (34.4%) resulted in the defendant being acquitted.
- If UCİM's request for participation in the case was denied, in 2 out of 12 cases (16.7%), the defendant received a sentence, and in 10 cases (83.3%), the defendant was acquitted.
- If UCİM's request for participation in the case was accepted, in 38 out of 49 cases (77.6%), the defendant received a sentence, and in 11 cases (22.4%), the defendant was acquitted.
- If UCİM's request for participation in the case was accepted and access to justice was timely, in 21 out of 22 cases

(95.5%), the defendant received a sentence, and in 1 case (4.5%), the defendant was acquitted.

- If UCİM's request for participation in the case was accepted but access to justice was not timely, in 17 out of 27 cases (63%), the defendant received a sentence, and in 10 cases (37%), the defendant was acquitted. If UCİM's request for participation in the case was accepted, access to justice was not timely, and the abuse was a one-time occurrence, in 8 out of 9 cases (88.9%), the defendant received a sentence.
- If UCİM's request for participation in the case was accepted, access to justice was not timely, and the abuse was systematic, in 9 out of 18 cases (50%), the defendant received a sentence, and in 9 cases (50%), the defendant was acquitted.
- If UCİM's request for participation in the case was accepted, access to justice was not timely, the abuse was systematic, and the type of abuse was non-qualitative sexual abuse, in 1 out of 7 cases (14.3%), the defendant received a sentence, and in 6 cases (85.7%), the defendant was acquitted.
- If UCİM's request for participation in the case was accepted, access to justice was not timely, the abuse was systematic, and the type of abuse was qualified sexual abuse, in 8 out of 11 cases (72.7%), the defendant received a sentence, and in 3 cases (27.3%), the defendant was acquitted.
- If UCİM's request for participation in the case was accepted, access to justice was not timely, the abuse was systematic, the type of abuse was qualified sexual abuse, and the child's age was below 11, in 2 out of 5 cases (40%), the defendant received a sentence, and in 3 cases (60%), the defendant was acquitted.
- If UCİM's request for participation in the case was accepted, access to justice was not timely, the abuse was systematic, the type of abuse was qualified sexual abuse, and the child's age was above 11, the defendant received a sentence in all 6 cases (100%).

In the subsequent section of this study, five independent variables (UCİM's participation in the case, the child's age, the type of abuse, the duration of the abuse, and timely access to justice) identified by the CHAID algorithm as predictors of the dependent variable will be examined using machine learning techniques, specifically logistic regression analysis and support vector machines, and the results will be compared.

The logistic regression method was used to assess whether the variables in the model were significant, and the significance of the independent variable coefficient estimates was tested at a 0.05 error rate. The classification matrix of the obtained logistic regression model was calculated using Table 5, and the accuracy value was determined to be 88.5%.

Table 5. Classification matrix of the logistic regression model

		Predicted	
Observed	Sentenced (Yes)	Not sentenced (No)	Percent correct
Sentenced (Yes)	38	11	0.775
Not sentenced (No)	2	10	0.833
Overall percentage	0.656	0.344	0.804

The effects of the five independent variables on the dependent variable were examined through logistic regression, and the results are presented in Table 6.

Table 6. Effects of five independent variables on the dependent variable through logistic regression analysis

	Estimate	**Std. Error**	**Z Value**	**Pr (>|z|)**	**Exp**
Intercept	-4.3184	1.9680	-2.194	0.02821*	0.0133
Child's Age2	-0.4729	1.1239	-0.421	0.67389	0.6232
Child's Age3	-1.8525	1.1378	-1.628	0.10350	0.1569
Recurrence/Duration of Abuse2	1.5555	1.0837	1.435	0.15117	4.7375
Type of Abuse2	1.2357	0.9206	1.342	0.17952	3.4407
Timely Access to Justice2	3.0111	1.0485	2.872	0.00408**	20.3096
UCİM's Participation in the Case2	3.4931	1.1178	3.125	0.00178**	32.8875
AIC	61.869				
Residual Deviance	47.869				

According to the logistic regression analysis, there are two independent variables that affect the dependent variable. These variables are UCİM's participation status in the case and the timely access to justice.

When testing the new model established with five independent variables that significantly affect the dependent variable, it was found that two of these variables were significant. The results of this new model, which includes these two independent variables, are presented in Table 7, and it was found that both variables are significant.

Table 7. Effects of five independent variables on the dependent variable through logistic regression analysis

	Estimate	Std. Error	Z Value	Pr (>|z|)	Exp
Intercept	-2.5661	0.7585	-3.383	0.000717***	0.0768
Timely Access to Justice2	1.9439	0.8243	2.358	0.018366*	6.9860
UCİM's Participation in the Case2	3.3479	1.0009	3.345	0.000823***	28.4439
AIC	61.707				
Residual Deviance	55.707				

In this model, the two independent variables that significantly affect the outcome of the case are UCİM's participation in the case and the timeliness of access to justice.

In cases where UCİM's participation request is accepted, the probability of the defendant receiving a sentence is 28.44 times higher compared to cases where the request is not accepted.

In cases where there is timely access to justice, the probability of the defendant receiving a sentence is 6.99 times higher compared to cases where there is a delay in access to justice.

In the support vector machines method, linear, radial, polynomial, and sigmoid kernel functions were examined separately for the 5 independent variables found to be significant in the CHAID analysis and the 2 independent variables found to be significant in the logistic regression. The results are compared in Table 8.

Table 8. Comparison of models based on 5 and 2 independent variables using support vector machines

Models Built with 5 Independent Variables	**Models Built with 2 Independent Variables**
In the linear model, the number of support vectors is 31 (14-17), Est. 0 1 0 38 7 1 2 14 and 52 out of 61 data points are correctly classified.	In the linear model, the number of support vectors is 28 (14-14), Est. 0 1 0 38 11 1 2 10 and 48 out of 61 data points are correctly classified.
In the radial model, the number of support vectors is 34 (17-17), Est. 0 1 0 38 5 1 2 6 and 54 out of 61 data points are correctly classified.	In the radial model, the number of support vectors is 30 (15-15), Est. 0 1 0 38 11 1 2 10 and 48 out of 61 data points are correctly classified.
In the polynomial model, the number of support vectors is 35 (17-18), Est. 0 1 0 38 8 1 2 13 and 51 out of 61 data points are correctly classified.	In the polynomial model, the number of support vectors is 30 (15-15), Est. 0 1 0 38 11 1 2 10 and 48 out of 61 data points are correctly classified.
In the sigmoid model, the number of support vectors is 33 (16-17), Est. 0 1 0 38 11 1 2 10 and 48 out of 61 data points are correctly classified.	In the sigmoid model, the number of support vectors is 31 (15-16), Est. 0 1 0 38 11 1 2 10 and 48 out of 61 data points are correctly classified.

In the support vector machines method using the radial basis function, which minimizes misclassification, the accuracy rates for the models with 5 and 2 independent variables are determined to be 88.5% and 78.7%, respectively. The metrics of the machine learning methods applied to the data subject to the research are compared in Table 9.

Table 9. Comparison of the metrics of machine learning techniques

Technique	Var.	AC	Sensitivity	Precision	F1
CHAID	5	0.885	0.875	0.946	0.91
CHAID	2	0.705	0.921	0.70	0.79
SVM	5	0.885	0.883	0.95	0.915

SVM	2	0.787	0.775	0.95	0.8536
Log. Reg.	5	0.787	0.775	0.95	0.8536
Log. Reg.	2	0.787	0.775	0.95	0.8536

As explained in Table 9, the techniques with the highest accuracy values are the models built using the CHAID algorithm and support vector machines with 5 independent variables. These models have also demonstrated significantly higher performance compared to other models based on their F1 scores.

4. Conclusion and Discussions

This study aims not only to identify the factors affecting the outcomes of child abuse cases but also to compare the performance of artificial intelligence and machine learning algorithms, which are rapidly advancing with modern technology, against traditional statistical methods. Additionally, it seeks to determine the hyperparameters and other characteristics of the model that delivers the highest performance.

When the CHAID algorithm was applied, the most influential independent variables affecting case outcomes were identified. The performance of models built using these variables with Decision Tree (DVM) and logistic regression algorithms was compared. It was found that the performances of the CHAID and DVM algorithms were equivalent and higher than that of logistic regression. The CHAID algorithm determined that the most influential independent variable was the "UCIM's involvement in the case.

One of the limitations of this study is the small number of cases (61) examined. It is quite difficult to access cases due to the lack of an open-access database containing child abuse lawsuit data in Turkey. Another limitation is the inability to compare the results of this study with previous research, as there are no studies on this subject.

It is recommended that new research be conducted by increasing the number of data points, expanding the number of cities or countries from which the data is collected, and diversifying the applied machine learning techniques. The machine learning techniques used in this study can be compared with traditional statistical methods. Additionally, different methods for model selection can be explored, and various model selection criteria can be tested. By comparing the results, the best-performing model can be identified.

Acknowledgments (if any)

This study was produced from the doctoral thesis of author, S. Şule Aksakal.

Funding

No funding was received for this work.

Credit authorship contribution statement

S. Şule Aksakal: Methodology, Software. **Erol Eğrioglu:** Methodology, Software, Writing, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

For the case information used in this study, necessary permissions were obtained from UCİM Saadet Öğretmen Association Struggling Child Abuse with the official letter dated 17.03.2022 and numbered 56-23032022-8. The permission document is provided as an appendix.

The information regarding child abuse cases used in this study is not publicly available in order to protect the personal rights and privacy of the children and their families involved as victims in the cases.

References

Acuna E, Rodriguez C, (2004). The Treatment of Missing Values and Its Effect on Classifier Accuracy. In: Classification, Clustering, and Data Mining Applications, Springer, Berlin, Heidelberg, 639-647.

Agresti, A., (2019). An Introduction to Categorical Data Analysis, 3rd ed. Wiley pp 156-190.

Akin, M., Eyduran, E. & Reed, B.M. (2017). Use of RSM and CHAID data mining algorithm for predicting mineral nutrition of hazelnut. Plant Cell Tiss Organ Cult 128, 303–316. https://doi.org/10.1007/s11240-016-1110-6

Allen-Collinson, J., (2009) A marked man: A case of female-perpetrated intimate partner abuse, International Journal of Men's Health, 8 (1): 22-40.

Alp, S. & Öz, E.,(2019). Classification Methods and R Applications in Machine Learning, *Nobel*, Ankara, 140-175.

Aydav, P.S.S., Minz, S., (2020). Granulation-based self-training for the semi-supervised classification of remote-sensing images. Granul. Comput. 5, 309–327 https://doi.org/10.1007/s41066-019-00161-x

Bajpai, A. (2018). *Child rights in India: Law, policy, and practice*. Oxford University Press.

Bowlby, J. (1984). Violence in the family as a disorder of the attachment and caregiving systems. *American journal of psychoanalysis*, *44*(1), 9.

Castro, F., Vellido, A., Nebot, À., Mugica, F. (2007). Applying Data Mining Techniques to e-Learning Problems. In: Jain, L.C., Tedman, R.A., Tedman, D.K. (eds) Evolution of Teaching and Learning Paradigms in Intelligent Environment. Studies in Computational Intelligence, vol 62. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-71974-8_8 pp 183-221

Chadaga, K., Prabhu, S., Sampathila, N. *et al.* (2024). An Explainable Framework to Predict Child Sexual Abuse Awareness in People Using Supervised Machine Learning Models. *J. technol. behav. sci.* **9**, 346–362. https://doi.org/10.1007/s41347-023-00343-0

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, 9, 13.

Chen, M.S., Han, J., Yu, P., (1996). "Data mining: an overview from a database perspective", *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 866-83.

Cheng, G., Chen, Q. & Zhang, R. (2021). Prediction of phosphorylation sites based on granular support vector machine. Granul. Comput. 6, 107–117. https://doi.org/10.1007/s41066-019-00202-5

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), 1178-1192.

Fan, MH., Chen, MY. & Liao, EC., (2021). A deep learning approach for financial market prediction: utilization of Google trends and keywords. Granul. Comput. 6, 207–216. https://doi.org/10.1007/s41066-019-00181-7

Fouché, G., Langit, L. (2011). Introduction to Data Mining. In: Foundations of SQL Server 2008 R2 Business Intelligence. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4302-3325-1_14 pp 369-402

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, *51*(1), 141-154.

Kass, G. (1980). An Exploratory Tecnique For İnvestigating Large Quantaties Of Categorical Data. Journal of the Royal Statistical Society., 29, 119-127.

Kassin, S. M., & Gudjonsson, G. H. (2004). The psychology of confessions: A review of the literature and issues. *Psychological science in the public interest*, *5*(2), 33-67.

Kruttschnitt, C., Kalsbeek, W. D., & House, C. C. (Eds.). (2014). *Estimating the incidence of rape and sexual assault* (pp. 48109-1382). Washington, DC: National Academies Press.

Lippard ETC, Nemeroff CB. (2020 Jan). The Devastating Clinical Consequences of Child Abuse and Neglect: Increased Disease Vulnerability and Poor Treatment Response in Mood Disorders. Am J Psychiatry. 1;177(1):20-36.

Liu, H., Cocea, M. (2019). Granular computing-based approach of rule learning for binary classification. Granul. Comput. 4, 275–283. https://doi.org/10.1007/s41066-018-0097-2

Maharana, K., Mondal, S., Nemade, B., (2022). A review: Data pre-processing and data augmentation techniques, Global Transitions Proceedings, Volume 3, Issue 1, June 2022, Pages 91-99 https://doi.org/10.1016/j.gltp.2022.04.020

Mathews, B., & Collin-Vézina, D. (2019). Child Sexual Abuse: Toward a Conceptual Model and Definition. Trauma, Violence, & Abuse, 20(2), 131-148. https://doi.org/10.1177/1524838017738726

Nguyen-Thihong, D., Vo-Van, T. (2024). Classifying for interval and applying for image based on the extracted texture feature. Granul. Comput. 9, 29 https://doi.org/10.1007/s41066-024-00450-0

Noble, W. (2006). What is a support vector machine?. *Nat Biotechnol* **24**, 1565–1567. https://doi.org/10.1038/nbt1206-1565

Paine, M. L., & Hansen, D. J. (2002). Factors influencing children to self-disclose sexual abuse. *Clinical psychology review*, *22*(2), 271-295.

Pant, M., Kumar, S. (2022). Fuzzy time series forecasting based on hesitant fuzzy sets, particle swarm optimization and support vector machine-based hybrid method. Granul. Comput. 7, 861–879. https://doi.org/10.1007/s41066-021-00300-3.

Raschka, S., Patterson, J., Nolet, C. (2020). Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. *Information* , *11*, 193. https://doi.org/10.3390/info11040193

Rybak, N., Hassall, M. (2022). Machine Learning–Enhanced Decision-Making. In: Hussain, C.M., Di Sia, P. (eds) Handbook of Smart Materials, Technologies, and Devices. Springer, Cham. https://doi.org/10.1007/978-3-030-84205-5_20

Tso, B. ve Mather P. M., (2009). Classification Methods For Remotely Sensed Data, Second Editon, Taylor & Francis Group, United States of America

Vapnik, V.(1963). Pattern recognition using generalized portrait method. Autom. Remote. Control. **24**, 774–780

Walker-Descartes, I., Hopgood, G., Condado, L. V., & Legano, L. (2021). Sexual violence against children. *Pediatric Clinics*, *68*(2), 427-436.

Wang, X., Ding, W., Liu, H. et al. (2020). Shape recognition through multi-level fusion of features and classifiers. Granul. Comput. 5, 437–448 https://doi.org/10.1007/s41066-019-00164-8

World Health Organization 2002. *World Report On Violence And Health: Summary*. Geneva

Yücesoy, E., Egrioglu, E. & Bas, E. (2023). A new intuitionistic fuzzy time series method based on the bagging of decision trees and principal component analysis. Granul. Comput. 8, 1925–1935 https://doi.org/10.1007/s41066-023-00416-8.

Zalcberg, S. (2017). The place of culture and religion in patterns of disclosure and reporting sexual abuse of males: A case study of ultra orthodox male victims. *Journal of Child Sexual Abuse*, *26*(5), 590-607.